Q: What is minimum % or number of reads that need to be present in a sample for a variant to be counted and included?

A: The variants are called by the individual studies with their own filtering and metrics. dbSNP only takes the final call sets and those that are marked as 'PASSED'. The advantage of ALFA is that it combined data across different studies so it should have reduced bias from a particular platform and filtering technique.

Q: I don't know much about coding, can I get the same results you got with APIs (E-Utilities, Variation Services) if I use the web pages?

A: Yes, the web can provide the records one at a time. Please take a look at the examples here for Entrez web searches https://www.ncbi.nlm.nih.gov/snp/docs/entrez_help/. The main advantage to using the APIs and the Python or other code is that you can retrieve information for large numbers of SNPs at once or from the entire release in an automated way.

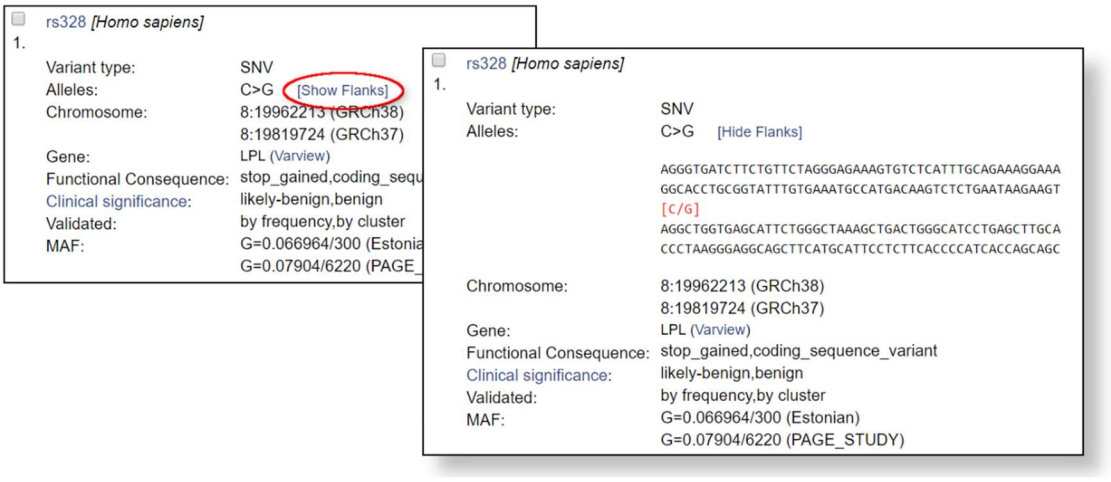
Q: The link to the supplement VCF FTP downloads is not in the slides. Would you post it?

A: The link is: https://ftp.ncbi.nlm.nih.gov/snp/population_frequency/latest_release/supplement/

Q: How can I get flanking sequence for a SNP?

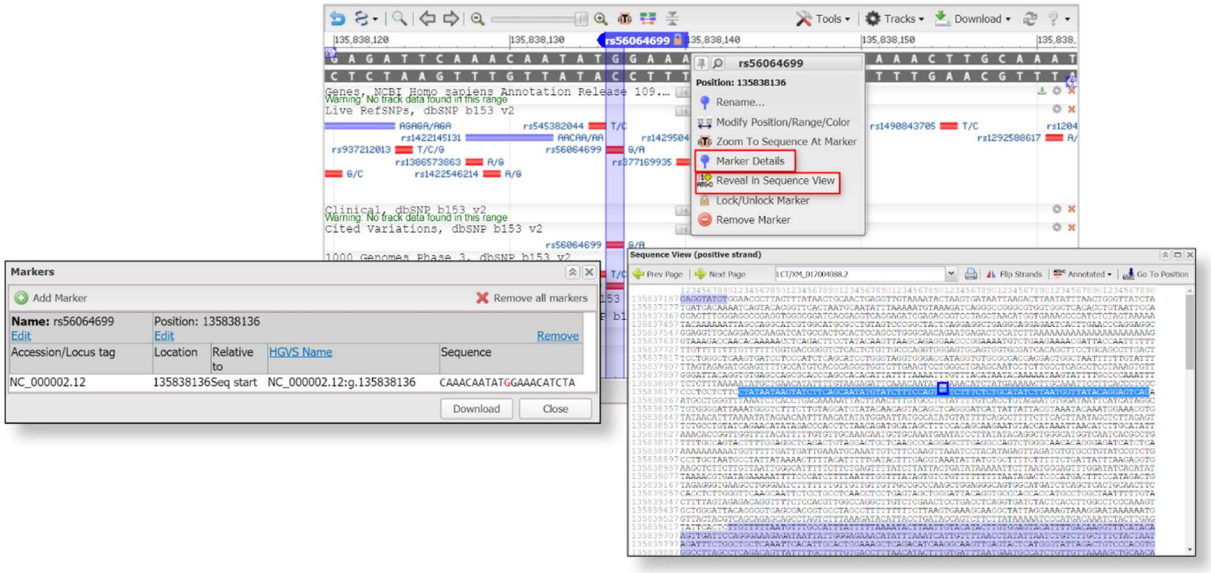
A: Here are three current ways to get the sequence flanking a SNP. We will also soon add a tab for 'FLANKS' on the RefSNP page.

1) The flanking sequence is available for a record on the web in SNP (https://www.ncbi.nlm.nih.gov/snp/?term=rs328) by clicking on the 'Show Flanks' next to the 'Alleles' description.



2) See the tutorial on GitHub to get flanking sequence programmatically using EUtils (https://github.com/ncbi/dbsnp/blob/master/tutorials/extract_flank.sh)

3) You can get the flanking sequences using the graphical sequence viewer at the bottom of a SNP page, for example https://www.ncbi.nlm.nih.gov/snp/rs56064699.
    a) Mouse over the marker near the lock icon and right click to bring up the pop-up menu options.
    b) Select “Marker Detail” from pop-up menu.
    c) Copy the flanking sequences in the marker detail box.
    d) If you want longer flanking sequences, select “Reveal in Sequence View” in Step #2 above.
    e) See the screenshot below for additional help.



Q: Can you run searches as a batch query?

A: You can use the three methods described below. The old dbSNP batch query service is discontinued.

1) You can use the E-Utilities or the Variation Services APIs as shown in the webinar. Use epost to upload dbSNP RefSNP (rs) numbers https://www.ncbi.nlm.nih.gov/books/NBK25500/. The UID for dbSNP is the rs number without the ‘rs’ prefix.

2) A tutorial to run a batch search using Variation Service is on GitHub https://github.com/ncbi/dbsnp/blob/master/tutorials/Variation%20Services/spdi_batch.py

3) Variation Services also provide a batch like request for VCF and HGVS formats. Each request can have up to 50,000 variants.

https://api.ncbi.nlm.nih.gov/variation/v0/#/HGVS/post_hgvs_batch_contextuals

https://api.ncbi.nlm.nih.gov/variation/v0/#/VCF/post_vcf_file_set_rsids

Q: Can you get results with deprecated rs numbers?

A: If you mean those that have merged with other SNP records, yes, a search with one of these will show you the record it merged with, for example https://www.ncbi.nlm.nih.gov/snp/?term=rs630496

Q: Why are 41% of ClinVar variants missing from ALFA?

A: Many ClinVar variants are rare and so may not show up with a measurable frequency in these samples. We may observe more ClinVar variants in future ALFA releases as more subjects are included.

Q: What attributes that refer to a specific allele/mutation (such as a SPDI representation, or a ClinGen Allele) can be searched in ALFA?

A: You can't search SNP directly with a ClinGen Allele or the SPDI representation. However, you can use the Variation Services API to convert the SPDI representation to the SNP identifier. You can also use Variation Services to remap the variant to GRCh38 and search ALFA. We'll investigate adding ClinGen Allele and SPDI notations to the Entrez web search in future releases.

Q: Do the samples used in the ALFA data overlap with the samples used in the genome aggregation database (gnomAD)?

A: Yes, it's possible some samples are shared across studies such as control sets or shared cohorts.

Q: Is it possible to split the allele frequencies by whole-exome, whole-genome?

A: Yes, but we don't have a plan to do so in the immediate future unless there is significant demand from the user community.

Q: If this data includes cancer patient data, does it mean this allele information still could contain cancer germline mutations?

A: Yes, it's possible but these mutations presumably are rare and do not occur broadly across all populations. You can set up your own filtering threshold using the ALFA allele frequency.

Q: Is it possible to get allele frequencies that exclude samples from patients with cancer?

A: Not in the initial release. ALFA provides a comprehensive catalog of all variants including common, rare, and possible pathogenic variants for broad use. You can set up your own filtering threshold using the ALFA allele frequency. dbSNP will consider providing flags for filtering somatic candidates in future releases.

Q: Do the allele frequencies include somatic variants or variants from tumor samples? Can we tell which ones those are or exclude them?

A: See the answer above.